心理科学进展 2019, Vol. 27, No. 5, 937–950
Advances in Psychological Science DOI: 10.3724/SP.J.1042.2019.00937

·研究方法(Research Method)·

解释性项目反应理论模型：理论与应用*

陈冠宇 陈 平
(北京师范大学中国基础教育质量监测协同创新中心，北京 100875)

摘 要 解释性项目反应理论模型(*Explanatory Item Response Theory Models,* EIRTM)是指基于广义线性混合模型和非线性混合模型构建的项目反应理论(*Item Response Theory,* IRT)模型。EIRTM 能在 IRT 模型的基础上直接加入预测变量，从而解决各类测量问题。首先介绍 EIRTM 的相关概念和参数估计方法，然后展示如何使用 EIRTM 处理题目位置效应、测验模式效应、题目功能差异、局部被试依赖和局部题目依赖，接着提供实例对 EIRTM 的使用进行说明，最后对 EIRTM 的不足之处和应用前景进行讨论。

关键词 解释性项目反应理论；广义线性混合模型；非线性混合模型；测量不变性；解释性测量
分类号 B841

1 引言

以 Binet 和 Simon (1904)的开创性工作为起点，项目反应理论(*Item Response Theory*, IRT)经过百余年发展，已广泛用于题目的标定与分析、被试的拟合与评分、测验的设计以及大规模教育评价等领域中(van der Linden, 2018)，是心理与教育测量领域最为重要的分析方法之一。虽然研究者针对作答评分、测验维度以及层级数据(*hierarchical data*)等实际问题提出一系列不同的模型并拓展 IRT 的应用情境，但是绝大部分 IRT 模型只能刻画被试与题目之间的关系，限制了 IRT 模型在心理与教育研究中的应用。

本文将基于广义线性混合模型(*Generalized Linear Mixed Models*, GLMM)和非线性混合模型(*Nonlinear Mixed Models*, NLMM)构建的 IRT 模型，定义为解释性项目反应理论模型(*Explanatory IRT Models,* EIRTM; De Boeck & Wilson, 2004)。EIRTM 是一个综合的解释性模型框架，它允许在 IRT 模型的基础上加入预测变量，在刻画被试和题目间关系的基础上，进一步解释相关变量影响，因而拓展 IRT 模型的应用范围。EIRTM 之所以重要，主要有以下几个方面的原因：

首先，EIRTM 摆脱传统 IRT 模型的限制，它不仅是测量模型，而且被称为解释性测量(*explanatory measurement*)模型。EIRTM 能够将题目特征和被试特征纳入模型并解释作答反应如何受到这些变量的影响，所以 EIRTM 可用于处理各种测量准确性问题：比如，题目位置效应(*Item Position Effect*, IPE)、测验模式效应(*Test Mode Effect*, TME)、题目功能差异(*Differential Item Functioning*, DIF)以及局部依赖(*Local Dependencies*, LD)等等。

其次，EIRTM 提出一个综合的模型构建观点。现有的 IRT 模型采用不同的术语标注和建模方法，使得研究者很难意识到 IRT 模型之间存在的共性(Rabe-Hesketh & Skrondal, 2016)。但是，绝大部分 IRT 模型实际上可以等价地构建为 GLMM 和 NLMM 的形式[1](De Boeck & Wilson, 2004, 2016; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003)。另外，EIRTM 体现 IRT 和回归分析的统一，是一个更为

收稿日期：2018-06-07
* 国家自然科学基金青年基金项目(31300862)，东北师范大学应用统计教育部重点实验室开放课题(KLAS130028732)和中国基础教育质量监测协同创新中心研究生自主课题(BJSM-2016A1-16004)资助。
通信作者：陈平, E-mail: pchen@bnu.edu.cn

[1] 不包括以三参数逻辑斯蒂克模型(Birnbaum, 1968)为代表的混合模型(*mixture models*)。

广义的分析框架。广义线性模型(*Generalized Linear Models*, GLM)涵盖以 *logit* 回归、*probit* 回归和基本线性模型(*basic linear models*)为代表的常用回归模型(Gill, 2000), 而且 GLM 和大部分 IRT 模型都是 GLMM 和 NLMM 的特例(Stroup, 2012)。因此通过引入 EIRTM 的框架, 研究者能够将回归模型和 IRT 模型涵盖在一个更为广义的分析框架之下, 从而形成更为完备的统计测量观。

最后, 应用 EIRTM 的最大优势在于对预测变量的直接建模和估计, 即“一步法”。虽然在实际应用中也可以采用“两步法”进行分析(即第一步先使用 IRT 模型得到不同测验情境[2]的参数估计值; 第二步再对不同情境得到的参数估计值进行显著性检验, 或者以参数估计值为因变量进行回归分析), 但是“一步法”要优于“两步法”: (1)“两步法”容易低估测量误差, 尤其是第一步分析中产生的测量误差经常会被忽视, 从而导致犯第一类错误的概率增大(刘红云, 骆方, 2008); (2) 相比于事先采用等组设计或事后采用多组比较的“两步法”, 采用“一步法”的 EIRTM 更为简便、也能处理更复杂的情况(Debeer & Janssen, 2013); (3) 使用 EIRTM 可将预测变量的效应与题目难度、被试能力分离, 这有助于对预测变量进行分析和解释(聂旭刚, 陈平, 张缨斌, 何引红, 2018)。

综上, EIRTM 提供一个灵活且综合的解释性模型框架。在 EIRTM 中, 研究者可以自主地构建研究所需要的 IRT 模型, 从而更好地解释数据。鉴于 EIRTM 的理论意义与应用价值, 本文将简单介绍 EIRTM 的基本理论并着重介绍 EIRTM 的应用情况, 以期能够帮助读者更加深入地了解和使用 EIRTM。本文将按以下顺序进行组织: 第 2 节概述 EIRTM 的基本概念以及参数估计方法; 第 3 节介绍如何使用 EIRTM 解决测量准确性问题; 第 4 节将提供一个具体例子对 EIRTM 的使用进行说明; 第 5 节讨论 EIRTM 的不足之处以及今后的研究方向。

2 EIRTM 的基本概念与模型参数估计

因为 GLMM 本质上是回归模型的拓展, 所以为了更好地理解 GLMM, 先简单引入线性回归模型(*linear regression model*):

$$Y_{pi} = \beta_0 + \beta_1 X_i + \varepsilon_{pi} \quad (1)$$

其中 p 代表被试, i 表示处理, β_0 为截距, β_1 为斜率, X_i 为预测变量的值, ε_{pi} 为残差。GLMM 是线性回归模型的一般形式。下面将具体介绍 GLMM 及 NLMM。

2.1 EIRTM 的基石: GLMM 和 NLMM

在预测变量与观测值建立连接之前使用连接函数(*link function*)进行转换的模型, 即 GLM。GLM 实际上就是经典回归模型的普遍化, 之所以称为“广义(*generalized*)”是因为连接函数可以任意选取。公式(1)所示的线性回归模型即用线性函数连接预测变量和观察值, 即本身连接函数(*identity link function*)。如果 GLM 中还包含随机效应(*random effect*), 那么模型就被称为 GLMM (Stroup, 2012)。随机效应是指预测变量的效应不是一个常数, 而是来源于一个概率分布, 具有期望和方差[3]; 与之对应的是固定效应(*fixed effect*), 是指预测变量的效应是一个常数, 没有测量误差[4]。在公式(1)中, 截距 β_0 和斜率 β_1 都是固定效应。

GLMM 由三个部分组成(De Boeck & Wilson, 2004):

(1) 随机成分(*random component*), 即观测变量及其期望的分布函数, 对应 IRT 中被试 p 在题目 i 上的作答反应 Y_{pi} 及其均值 μ_{pi} 的分布函数。当作答反应为二分时, 其分布函数为独立的伯努利分布(*Bernoulli distribution*), 记为 $Y_{pi} \sim \text{Bernoulli}(\pi_{pi})$, 其中 π_{pi} 表示被试 p 在题目 i 上的正确作答概率 $P\left(Y_{pi}=1\right)$ 且 $\mu_{pi} = \pi_{pi}$。

(2) 连接函数, 即用于连接观测变量的期望 π_{pi} 和系统成分 η_{pi}, 记为 $\eta_{pi} = f_{link}\left(\pi_{pi}\right)$, 其中 $f_{link}(\cdot)$ 表示连接函数。在 IRT 领域中, 可以使用 *probit* 连接函数和 *logit* 连接函数, 它们分别对应正态肩形模型(*normal-ogive models*)和逻辑斯蒂克

[2] 不同的测验情境是指不同的题本、不同的被试群体或者不同的测验形式等等, 本质上就是 IRT 研究中的多组分析(*multiple group analysis*)。

[3] 在 IRT 模型中引入随机效应看似不常见, 但 EM 算法的最大边际似然估计(*Maximum Marginal Likelihood Estimation with EM*, MMLE/EM)就是将伴随参数(*incidental parameter*, 即能力参数)视为随机效应(Bock & Aitkin, 1981; Bock & Lieberman, 1970)。

[4] 这些概念经常用于多层线性模型(*Hierarchical Linear Model*, HLM)中。本质上, 随机效应对应的随机系数回归方法(*random coefficients approach*)也被称为分层回归方法或多水平回归方法(*hierarchical or multilevel regression approach*)。

模型(*logistic models*)。

(3) 系统成分(*systematic component*)，即预测变量的线性函数，记为 η_{pi}。在 GLMM 中，预测变量可以分为两类，具有固定效应 β_q 的预测变量 X_{iq} 和具有随机效应 θ_{pj} 的预测变量 Z_{ij}：

$$\eta_{pi} = logit(\pi_{pi}) = \log\left(\frac{P(Y_{pi}=1)}{1-P(Y_{pi}=1)}\right) = \sum_{j=1}^{J}\theta_{pj}Z_{ij} - \sum_{q=1}^{Q}\beta_q X_{iq} \qquad (2)^{5}$$

其中 i 对应题目，p 对应被试；Q 和 J 分别表示固定效应 β_q 和随机效应 θ_{pj} 的个数，X_{iq} 和 Z_{ij} 为预测变量。此处假设 X_{iq} 为题目的指示变量(*indicator variable*)，即题目的虚拟编码(*dummy code*)变量，当 $i=q$ 时，$X_{iq}=1$，当 $i \neq q$，$X_{iq}=0$；Z_{ij} 同理，也可视为维度的指示变量。记 $\boldsymbol{\theta}_p=(\theta_{p1},\theta_{p2},\cdots,\theta_{pJ})^T$，有 $\boldsymbol{\theta}_p \sim N(0,\boldsymbol{\Sigma})$，即 $\boldsymbol{\theta}_p$ 服从均值向量为 0、协方差矩阵为 $\boldsymbol{\Sigma}$ 的多元正态分布[6]。在 GLMM 中，η_{pi} 只由线性成分构成，对应 Rasch 模型簇。但是对于包含区分度参数的 IRT 模型来说，还包括非线性成分(参数相乘)，属于 NLMM[7]。因此，通过 GLMM 和 NLMM 构建 EIRTM，就能从更一般的视角拓展 IRT 模型，详见第 4 节的 EIRTM 实例部分。

2.2　EIRTM 的参数估计

EIRTM 的参数估计方法有很多，但都涉及复杂的统计知识，此处仅做简单介绍：(1)全似然分析(*full-likelihood analysis*)，即对 EIRTM 的边际似然函数进行数值逼近(*numerical approximation*)以求得估计值使边际似然函数达到最大值。此类方法包括高斯–厄尔米特求积(*Gauss-Hermite quadrature*)与蒙特卡罗积分(*Monte Carlo integration*)等直接最大法[对应的统计软件(包)为 SAS PROC NLMIXED (SAS Institute, 2015)、STATA 的 GLLAMM (Rabe-Hesketh, Skrondal, & Pickles, 2004)和 HLM (Raudenbush, Bryk, Cheong, Congdon Jr, & Toit, 2011)]以及使用 EM 算法的间接最大法[对应的软件有 MULTILOG (Thissen, 1991)和 ConQuest (Adams, Wu, & Wilson, 1988)]；(2)线性分析近似(*linearized analytical approximations*)，即对 EIRTM 的边际似然函数中含有的积分求近似解，包括拉普拉斯近似(*Laplace approximation*)、带惩罚的拟似然法(*Penalized Quasi-Likelihood Method*, PQL)和边际拟似然法(*Marginal Quasi-Likelihood Approach*, MQL)，对应的软件(包)有 R 语言的 lme4 包(Bates, Mächler, Bolker, & Walker, 2015)、HLM 和 SAS PROC GLIMMIX (SAS Institute, 2015)；(3)贝叶斯方法，即采用马尔科夫链蒙特卡洛 (*Markov chain Monte Carlo*, MCMC)方法，典型的分析软件有 OpenBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2014)。更详细的算法介绍与比较可以参见 Bolker 等(2009)的综述。

目前尚未发现不同方法得到的估计结果之间会存在显著差异。De Boeck 和 Wilson (2004)对 6 种统计软件的估计结果进行比较，发现差异不大，而且采用同一类估计方法的软件的估计结果更加接近。Jeon, Rijmen 和 Rabe-Hesketh (2013)基于模拟数据对 WinBUGS[8]、PROC NLMIXED、GLLAMM 以及含逻辑斯蒂回归节点的贝叶斯网络(*Bayesian Networks with Logistic Regression Nodes*, BNL; Rijmen, 2006)进行比较，结果发现：不同软件估计的结果相似，差别在于 BNL 的估计速度远快于其他软件。另外，Jeon, Rijmen 和 Rabe-Hesketh (2014)还在 BNL 的基础上，开发了 R 语言的 FLIRT 包。总之，目前用于分析 EIRTM 的软件种类繁多，但是

[5] 公式(2)是基于 IRT 模型改写的：(1) 此处 $\sum_{q=1}^{Q}\beta_q X_{iq}$ 对应题目 i 的难度 β_i ($\sum_{q=1}^{Q}\beta_q X_{iq}=\beta_i$)，即 $\beta_q=\beta_i$。此表达没有截距，也就是忽略 β_q 的均值 β_0；(2) 另一种常见写法是 $\sum_{j=1}^{J}\theta_{pj}Z_{ij}+\sum_{q=1}^{Q}\beta'_q X_{iq}$，其中的 $\sum_{q=1}^{Q}\beta'_q X_{iq}$ 可以理解为题目容易度(*item easiness*)；(3) 还有一种写法是将第一题作为参照题，截距为 β_0，下标从 0 开始直到 q-1 结束，而且 $\beta_0+\beta_{q-1}=\beta_i$，这种写法多用于多水平 IRT 模型。

[6] 据此，公式(2)可以表示成更简洁的矩阵形式：$\boldsymbol{\eta}_p = \boldsymbol{X\beta}+\boldsymbol{Z\theta}_p$。虽然矩阵形式在统计领域更为常见，但考虑到解释的便利和研究的实际，本文统一使用指示变量(虚拟变量)组织公式。

[7] 其实也可以说，GLMM 是 NLMM 的特例(Rijmen et al, 2003)，因为 NLMM 既能刻画非线性关系又能描述线性关系。

[8] 上文所述的 OpenBUGS 是 WinBUGS 的后续开源版本，两者几乎相同，详见 https://www.mrc-bsu.cam.ac.uk/software/bugs/。

不同软件估计结果接近，研究者可以根据自己的需要进行选择。

3 使用 EIRTM 处理测量准确性问题

3.1 题目位置效应(Item Position Effect, IPE)

IPE 是指同一个题目在不同测验间因题目位置的变化而导致题目参数的变化(聂旭刚等人, 2018)。IPE 违背了 IRT 的参数不变性(*parameter invariance*)特征，使得基于 IRT 的测验公平性分析、计算机化自适应测验(*Computerized Adaptive Testing*, CAT)以及矩阵抽样设计(*matrix sampling design*)等重要应用都受到影响。因此，很有必要对 IPE 进行检测及解释。

用于检测 IPE 的 EIRTM 可以分为三类(聂旭刚等人, 2018)：第 1 类模型记为模型 IPE-1 (Hohensinn, Kubinger, Reif, Schleich, & Khorramdel, 2011)：

$$\eta_{pi} = \theta_{p1} - \left[\sum_{q=1}^{Q} \beta_q X_{iq} + \gamma(k-1)\right] \qquad (3)$$

其中 p 表示被试，i 表示题目($i=1,2,\cdots,I$)，q 表示变量($q=1,2,\cdots,Q$)，且 $Q = I$；θ_{p1} 为能力参数，$\theta_{p1} \sim N\left(0,\sigma^2_{\theta_{p1}}\right)$；$X_{iq}$ 为指示变量，当 $i=q$ 时，$X_{iq}=1$，否则取 0；$\sum_{q=1}^{Q} \beta_q X_{iq}$ 如前文所述，对应题目难度；γ 表示的是 IPE。此时 γ 为固定效应，它只与题目位置 k 有关，所有题目在同一位置的难度变化都相同[9]。此模型本质上是对题目难度进行分解，从而得出 IPE。

第 2 类模型记为模型 IPE-2 (Debeer & Janssen, 2013)：

$$\eta_{pi} = \theta_{p1} - \left[\sum_{q=1}^{Q} \beta_q X_{iq} + \gamma_i(k-1)\right] \qquad (4)$$

注意此处 $\gamma_i = \gamma + \gamma_i', \gamma_i'$ 被定义为随机效应，$\gamma_i' \sim N\left(0,\sigma^2_{\gamma_i'}\right)$，其余参数含义同上。此模型假设 IPE 受题目的影响，即不同题目在同一位置上的难度变化不同。

第 3 类模型记为 IPE-3 (Hartig & Buchholz, 2012)：

$$\eta_{pi} = \theta_{p1} + \theta_{pk}(k-1) - \sum_{q=1}^{Q} \beta_q X_{iq} \qquad (5)$$

其中 θ_{pk} 是随机效应，$\theta_{pk} \sim N\left(0,\sigma^2_{\theta_{pk}}\right)$，表示 IPE。此时，IPE 可以被视为一个新的维度，有研究者将它解释为毅力(*persistence*)或考生努力(*examinee effort*; Debeer, Buchholz, Hartig, & Janssen, 2014)。此模型假设 IPE 与被试有关，即不同位置的题目难度受到被试的影响(Weirich, Hecht, Penk, Roppelt, & Böhme, 2017)。Debeer 和 Janssen (2013)对上述三类模型进行比较后认为第三类模型更有优势，即将 IPE 解释为被试层面的属性更符合实际。

IPE-1 假设 γ 由题目难度分解得到，而且不同题目的 γ 相同。本质上，γ 是预测变量 X_{i0} 的固定效应：X_{i0} 对于所有题目都取 1，γ 就是所有题目 IPE 的均值。IPE-2 加入的 γ_i 是基于题目的随机效应，表示不同题目的 IPE 可以不同。IPE-3 加入的 θ_{pk}，则是基于被试的随机效应，它表示不同被试的 IPE 可以不同。其实，固定效应和随机效应的选择完全基于研究者的需要，类似于“HLM 中设定斜率和截距是固定还是随机”。如果研究者认为 IPE 具有跨题目一致性，就可将 IPE 设定为固定效应；如果 IPE 在不同题目上不同，则可以用一个概率分布(随机效应)来表示 IPE。所以在 EIRTM 中，设定效应为固定或随机是非常灵活的：通常作为固定效应处理的题目也可以视为随机效应(De Boeck et al., 2011)，这等于带误差项的线性逻辑斯蒂克测验模型(*Linear Logistic Test Models*, LLTM; Janssen, 2016; Weirich, Hecht, & Böhme, 2014)。

3.2 测验模式效应(*Test Mode Effect*, TME)

国际大规模测评项目正在经历由纸笔测验(*Paper-Based Assessment*, PBA)形式向计算机化测验(*Computer-Based Assessment*, CBA)形式的转变。在国际学生能力评估项目(*Programme for International Student Assessment*, PISA) 2015 的技术报告中(OECD, 2017a)将 TME 定义为：被试在一种测验模式(如 PBA)中的表现与在同一个测验的另一种测验模式(如 CBA)中的表现相比，出现的功能性差异。TME 反映的是同一测验在不同测验模式下的结果不可比问题，它本质上是对测量不变性(*measurement invariance*)的研究。

为探究 TME 的实际影响，PISA 2015 使用了 3 个 EIRTM 模型，模型 1 记为 TME-1：

[9] 此处仅假设 IPE 为线性变化，更复杂的非线性情况可以表示为 k 的二次函数等(参见 Kang, 2014; Trendtel & Robitzsch, 2018)

$$\eta_{pim}=\left(\sum_{q=1}^{Q}\alpha_q X_{iq}\right)\left(\theta_{p1}-\sum_{q=1}^{Q}\beta_q X_{iq}+\delta_m M_{\{i>I\}}\right) \quad (6)$$

其中 i 代表题目($i=1,2,\cdots,2I$)，当 $i=1,\cdots,I$ 时，表示的是 PBA 中的题目，当 $i=I+1,I+2,\cdots,2I$ 时，表示的是与前 I 道题相同的题目，只是测验形式变成 CBA；q 表示变量(q = 1,2,L,Q, Q = 2I)；$M_{\{i>I\}}$ 是指示变量，当 $i>I$ 时 $M_{\{i>I\}}=1$，否则取 0，即 $M_{\{i>I\}}$ 是不同测验模式的虚拟编码变量；m 表示模式，δ_m 即 TME；$\sum_{q=1}^{Q}\alpha_q X_{iq}=\alpha_i$ 如前文所述，表示题目区分度；其余参数含义同上。假设 $i'\in\{I+1,I+2,\cdots,2I\}$，于是根据模型有 $\beta_{i'}=\beta_{i'-I}-\delta_m$，且假设 $\alpha_{i'-I}=\alpha_{i'}$。此模型表示任意 PBA 中的题目转换为 CBA 形式后，题目难度都受到相同的 TME (δ_m)影响，但题目区分度不受影响。

第 2 个模型记为 TME-2：

$$\eta_{pim}=\left(\sum_{q=1}^{Q}\alpha_q X_{iq}\right)\left(\theta_{p1}-\sum_{q=1}^{Q}\beta_q X_{iq}+\sum_{q=1}^{Q}\delta_{mi} M_{\{i>I\}} X_{iq}\right) \quad (7)$$

其中 δ_m 变为 δ_{mi}，对于某些题目而言，δ_{mi} 可能为 0，即不同测验模式的难度不变，不存在 TME；有些题目的 δ_{mi} 则不为零，即存在 TME。其余参数含义同上。对于前 I 道题目而言，因为 $M_{\{i>I\}}=0$，所以 $\sum_{q=1}^{Q}\delta_{mi} M_{\{i>I\}} X_{iq}=0$，于是前 I 道题目中的题目 j 的线性成分为 $\eta_{pjm}=\alpha_j\left(\theta_{p1}-\beta_j\right)$；对于后 I 道题目而言，因为 $M_{\{i>I\}}=1$，所以其中题目 j 的线性成分为 $\eta_{pjm}=\alpha_j\left(\theta_{p1}-\beta_j+\delta_{mj}\right)$。此模型假设 PBA 中的题目转换为 CBA 形式后，不同题目具有不同的 TME。

第 3 个模型记为 TME-3：

$$\eta_{pim}=\left(\sum_{q=1}^{Q}\alpha_q X_{iq}\right)\left(\theta_{p1}-\sum_{q=1}^{Q}\beta_q X_{iq}\right)+\sum_{q=1}^{Q}\alpha_{mi}\theta_{p2}M_{\{i>I\}}X_{iq} \quad (8)$$

其中 α_{mi} 是另一个斜率参数，称为模式斜率(*mode slope*)，反映被试的 TME 在不同题目上的影响不同；θ_{p2} 是另一个潜变量，表示 TME，为随机效应。假设两个随机效应不相关，即 $\mathrm{cov}\left(\theta_{p1},\theta_{p2}\right)=0$。类似地，对于前 I 道题目而言，其中题目 j 的线性成分为 $\eta_{pjm}=\alpha_j\left(\theta_{p1}-\beta_j\right)$；对于后 I 道题目而言，其中题目 j 的线性成分为 $\eta_{pjm}=\alpha_j\left(\theta_{p1}-\beta_j\right)+\alpha_{mj}\theta_{p2}$。此模型假设 TME 是基于被试的效应，也即不同被试具有不同的 TME。

综上，TME-1 和 TME-2 采用基于题目的固定效应(δ_m 和 δ_{mi})表示 TME，而 TME-3 则使用基于被试的随机效应(θ_{p2})表示 TME。如果认为 $M_{\{i>I\}}$ 是不同测验模式的分组变量，那么可以更准确地将 θ_{p2} 定义为被试和模式交互的随机效应。与 IPE 模型相比，建构 TME 模型的思路非常类似：IPE-1 和 TME-1 都加入一个跨题目一致的固定效应；而 IPE-2 和 TME-2 都是从题目的角度出发，认为效应跨题目不一致性，只不过 IPE-2 定义的效应是随机效应，而 TME-2 定义的是固定效应；IPE-3 和 TME-3 则都是从被试的角度出发，认为模型都受到基于被试的随机效应的影响。

PISA 采用真实数据对上述三个模型进行比较，结果发现：TME-3 的相对拟合指标最好，TME-2 的结果接近 TME-3，TME-1 的拟合最差；综合考虑模型的复杂性和数据拟合情况，TME-2 的表现最优。基于 TME-2 的结果还有：绝大多数的题目满足强测量不变性(*strong measurement invariance*)，即斜率和难度参数在不同测验模式下不变；部分题目满足弱测量不变性(*weak measurement invariance*)，即斜率参数不变、难度参数发生变化。可见，CBA 的使用确实会对评估学生成绩造成影响(Cosgrove & Cartwright, 2014; Logan, 2015)。值得注意的是，Jerrim (2016)发现中国上海的学生在 PISA 2015 出现显著的成绩降低，并且原因很可能就是 CBA 的使用。无独有偶，新西兰教育研究委员会(*New Zealand Council for Educational Research*, NZCER)对 PBA 和 CBA 进行比较，也发现学生成绩出现显著下降(Eyre, Berg, Mazengarb, & Lawes, 2017)。总之，TME 的存在已被证实，考虑 TME 相比不考虑修正 TME 能够更好地提升测验质量(Jerrim, Micklewright, Heine, Salzer, & McKeown, 2018)。

3.3　题目功能差异(Differential Item Functioning, DIF)

DIF 是指具有相同能力的被试(组)在作答相同题目时出现的功能性差异，这种差异是由被试

所处群体的不同而造成的。DIF 也属于测量不变性问题，反映的是题目受到与测验无关因素的影响。

用于 DIF 分析的 EIRTM 描述如下，记为 DIF-1 (De Boeck et al., 2011)：

$$\eta_{pi} = \theta_{p1} - \sum_{q=1}^{Q} \beta_q X_{iq} + \zeta_{focal} Z_g + \sum_{q=1}^{Q} \delta_{ig} X_{iq} Z_g \quad (9)$$

其中 ζ_{focal} 是目标组(*focal group*)和参照组(*reference group*)的总效应，也即两组被试能力均值之差; g 表示组，Z_g 是被试组别的指示变量，当被试 p 属于参照组时，$Z_g = 0$，当被试 p 属于目标组时，$Z_g = 1$; δ_{ig} 即题目 i 上 DIF 的效应量，δ_{ig} 本质上是被试组别和题目的交互，而且 δ_{ig} 只存在于目标组作答的题目 i 上，因为这时 $X_{iq} = 1$ 且 $Z_g = 1$；其余参数含义不变。当被试 p 属于目标组时，题目 j 的线性成分为：$\eta_{pi} = \theta_{p1} - \beta_j + \zeta_{focal} + \delta_{jg}$；当被试 p 属于参照组时，题目 j 的线性成分为：$\eta_{pi} = \theta_{p1} - \beta_j$。

注意此模型同时加入两个固定效应：(1) ζ_{focal} 用于控制目标组和参照组的能力均值差异，即被试群体间的真实能力差异，Osterlind 和 Evenson (2009)称之为“影响(*impact*)”。由于 ζ_{focal} 基于被试的组别得到，所以它是基于被试的固定效应。如果有证据支持两组之间没有能力差异或者已经通过匹配等手段进行控制，则可以移除此效应；(2) δ_{ig} 是被试组别和题目交互的固定效应，反映题目难度在组别上的变化。公式(12)假定参照组中所有题目都可能存在 DIF (通过指示变量 X_{iq} 定义)，实际上也可以自定义需要估计 DIF 的题目(如果不需要估计题目 j 的 DIF，则从 $\sum_{q=1}^{Q} \delta_{ig} X_{iq} Z_p$ 中移除含 X_{ij} 的项即可)。如何选取需要估计 DIF 的题目以及是否需要将有 DIF 嫌疑的题目从匹配标准中排除，则属于纯化(*purification*)的问题。

一些研究者基于贝叶斯方法估计 DIF-1 模型，因此称之为整合的贝叶斯 DIF 模型(*Integrated Bayesian DIF models*, IBDM)，IBDM 的估计结果优于传统的 DIF 方法(Gamerman, Gonçalves, & Soares, 2018)。还有研究将此类 DIF 模型应用于不同的情景和算法中，侦测出不同组别之间的 DIF 效应(Bechger & Maris, 2015; Tutz & Berger, 2016; Tutz & Schauberger, 2015)。总之，虽然此类 DIF 模型的应用情境有所不同，但是 DIF-1 模型最大的优势就是能够自由估计来自不同组别(协变量)的 DIF 效应。

3.4 局部依赖(*Local Dependence*, LD)

局部独立性(*Local Independence*, LI)是 IRT 理论的基本假设之一，与 LI 对立的概念是 LD。LD 可分为局部被试依赖性(*Local Person Dependence*, LPD)和局部题目依赖性(*Local Item Dependence*, LID)。LPD 是指在给定被试能力时，被试在不同题目的作答反应之间存在相依性；LID 指题目参数已知时，不同能力的被试在该题目上的作答反应间存在相依性(詹沛达, 王文中, 王立君, 2013)。

在 IRT 领域中，LPD 出现的主要原因是被试群组效应(*Person Clustering Effect,* PCE)。选取的被试嵌套于不同的群体，属于同一群体的被试可能受到相同的外部支持或干扰、具有同样的学习机会和采用相同的解题策略，因而有理由认为他们的作答相似，即存在 PCE (Jiao, Kamata, Wang, & Jin, 2012)。PCE 的存在使得样本量的影响变小，从而导致有偏的参数估计。为处理 PCE 导致的 LPD, Kamata (2001)提出三水平 IRT 模型，对应的层级关系如图 1 所示。在 EIRTM 框架下进行重新公式化后，可以得到 LPD-1：

$$\eta_{pi} = \theta_{p1} + \sum_{q=0}^{Q-1} \beta_q X_{iq} + \varepsilon_{pg} \quad (10)$$ [10]

其中 $\sum_{q=0}^{Q-1} \beta_q X_{iq}$ 较之前的表达略有改变，这表示以某一道题为参照题(一般取最后一题)，得到题目截距 β_0, β_1 即为题目 1 与参照题的难度之差，其余以此类推；故 X_{i0} 作为题目截距的指示变量，

[10] 原始公式基于多层广义线性模型(*Hierarchical Generalized Linear Model*, HGLM)，对 GLMM 增加限制条件就能得到 HGLM (De Boeck & Wilson, 2004)。此处保留了 HGLM 使用“+”连接被试和题目参数(此时 $\sum_{q=0}^{Q-1} \beta_q X_{iq}$ 解释为题目容易度)，并使用其中一个题目作为参照(故下标从 0 开始，$Q-1$ 结束)的习惯。此外，用 ε_{pg} 替换了文献中表示 PCE 的 ω_{00g}。这样处理的目的是希望读者能够理解 EIRTM 框架和 HGLM 的共性和符号注释上的细微差异。由于 HGLM 从属于GLMM的框架，也就是说多水平 IRT模型(*Multilevel Item Response Theory Model*)都可通过 EIRTM 构建。

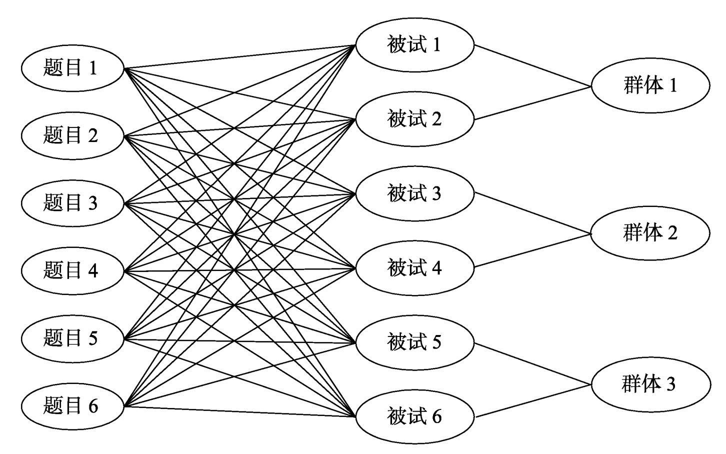


图 1　题目、被试和群体的层级关系图

注：图片翻译自 Jiao, Kamata 和 Xie (2015, p. 145) 图 5.3

取值固定为 1，其余 $\sum_{q=1}^{Q-1} X_{iq}$ 含义不变。ε_{pg} 表示的是被试 p 在群体 g 中的 PCE，为随机效应，$\varepsilon_{pg} \sim N\left(0,\sigma^2_{\varepsilon_{pg}}\right)$；其余参数含义不变。于是，被试 p 在题目 j ($j \neq I$)上的线性成分为：$\eta_{pj} = \theta_{p1} + \beta_0 + \beta_j + \varepsilon_{pg}$ (注意最后一题 I 上的线性成分为 $\eta_{pI} = \theta_{p1} + \beta_0 + \varepsilon_{pg}$)。此模型表示被试受到所属群体 PCE 的影响，而且同一群体中的被试受到的 PCE 相同。

在 IRT 领域中, LID 出现的主要原因是题组效应(*testlet effect*, TE)。题组是一组共用相同刺激材料的题目(Wang & Wilson, 2005)，因此被试对同一题组中不同题目的作答不再 LI，而存在 TE。忽视 TE 会对测验信度、被试能力、题目难度、题目区分度参数以及 DIF 分析造成影响(Bolt, 2002; Ip, 2000; Lee, 2004; Wainer & Lukhele, 1997; Wainer, Sireci, & Thissen, 1991)。包含 TE 的 IRT 模型如图 2 的右侧三列所示，记为 LID-1 (Jiao, Wang, & Kamata, 2005)：

$$\eta_{pi} = \theta_{p1} + \sum_{q=0}^{Q-1} \beta_q X_{iq} + \sum_{d=1}^{D} \gamma_{pd} T_{id} \tag{11}$$

其中 $\sum_{q=0}^{Q-1} \beta_q X_{iq}$ 同式(10); d 表示题组($d = 1,2,\cdots,D$); 引入指示变量 T_{id}，当题目 i 属于题组 d 时，$T_{id} = 1$，否则 $T_{id} = 0$; γ_{pd} 表示被试 p 在题组 d 中的 TE，γ_{pd} 是随机效应，有 $\gamma_{pd} \sim N\left(0,\sigma^2_{\gamma_{pd}}\right)$; $\sum_{d=1}^{D} \gamma_{pd} T_{id}$ 可以表示特定题目上的 TE；其余参数含义同上。假设题目 j ($j \neq I$)属于题组 1，题目 k ($k \neq I$)属于题组 2，对被试 p 有：$\eta_{pj} = \theta_{p1} + \beta_0 + \beta_j + \gamma_{p1}, \eta_{pk} = \theta_{p1} + \beta_0 + \beta_k + \gamma_{p2}$。可见通过使用 T_{id}，研究者可以在

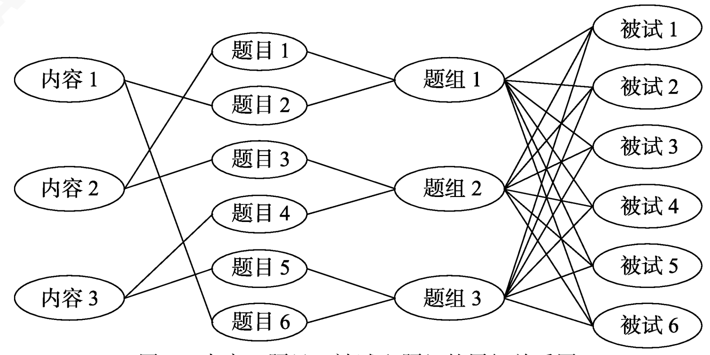


图 2　内容、题目、被试和题组的层级关系图

注：图片翻译自 Jiao 等(2015, p. 148) 图 5.5

EIRTM 中灵活定义测验的结构：无论是所有题目都基于题组构建，还是只有部分题目基于题组构建。此模型表示 TE 是基于被试的随机效应，即不同被试的 TE 存在差异。

此外，造成 LID 的原因还有可能是不同题目采用相同的测验内容，即存在内容群组效应(*Content Clustering Effect*, CCE)。因此，如图 2 所示，题目可以视为既嵌套于题组又嵌套于内容，即交叉分类(*cross-classified*)。考虑到此时有两个造成 LID 的因素，可称为双重 (*dual*) LID，将此模型记为 LID-2 (Xie, 2014; Xie & Jiao, 2014)：

$$\eta_{pi} = \theta_{p1} + \sum_{q=0}^{Q-1}\beta_q X_{iq} + \sum_{d=1}^{D}\gamma_{pd}T_{id} + \sum_{c=1}^{C}\gamma'_{pc}T'_{ic} \quad (12)$$

其中 $\sum_{q=0}^{Q-1}\beta_q X_{iq}$ 和 $\sum_{d=1}^{D}\gamma_{pd}T_{id}$ 同式(14); c 表示内容($c = 1,2,\cdots,C$); 引入指示变量 T'_{ic}，当题目 i 属于内容 c，$T'_{ic} = 1$，否则 $T'_{ic} = 0$; γ'_{pc} 表示被试 p 在内容 c 上的 CCE，γ'_{pc} 是随机效应，有 $\gamma'_{pc} \sim N\left(0, \sigma^2_{\gamma'_{pc}}\right)$；其余参数含义不变。同样地，也可以使用 T'_{ic} 灵活定义测验的内容结构。假设题目 j ($j \neq I$)属于题组 1 且属于内容 1，于是被试 p 在 j ($j \neq I$)上的线性成分为：$\eta_{pj} = \theta_{p1} + \beta_0 + \beta_j + \gamma_{p1} + \gamma'_{p1}$。在此模型中，CCE 和 TE 都是基于被试的随机效应，不同被试间可以存在差异。

最后，还可以将 LPD 和 LID 相结合，即在图 2 右侧的被试上再加入群体，从而构成最完整的 LD 模型，记为 LD-1 (Jiao et al., 2015)：

$$\eta_{pi} = \theta_{p1} + \sum_{q=0}^{Q-1}\beta_q X_{iq} + \sum_{d=1}^{D}\gamma_{pd}T_{id} + \sum_{c=1}^{C}\gamma'_{pc}T'_{ic} + \varepsilon_{pg} \quad (13)$$

其中的参数含义同上。假设题目 j 属于题组 1 且属于内容 1，于是被试 p 在 j ($j \neq I$)上的线性成分为：$\eta_{pj} = \theta_{p1} + \beta_0 + \beta_j + \gamma_{p1} + \gamma'_{p1} + \varepsilon_{pg}$。$\varepsilon_{pg}$ 的表示与 γ_{pd} 和 γ'_{pc} 略有不同，这是因为 PCE 与 TE、CCE 不属于同一个水平(层次)：(1) 对于 PCE 而言，一个合理的抽样设计不会出现“某些被试属于特定群体，而另外一些被试不属于任何群体”的情况，这样本身就会造成被试的异质性；(2) 对于 TE 和 CCE 而言，一个被试可能受到多个 TE 和 CCE 的影响，因此需要通过引入指示变量 T_{id} 和 T'_{ic} 来表示某个题目上的作答是否受到 TE 和 CCE 的影响以及受到哪个题组或内容的影响。当然，若整个测验只涉及一个题组和一个内容，那么 LD-1 可以简化为：$\eta_{pi} = \theta_{p1} + \sum_{q=0}^{Q-1}\beta_q X_{iq} + \gamma_{pd}T_{id} + \gamma'_{pc}T'_{ic} + \varepsilon_{pg}$。

Jiao 等人(2015)基于 PISA 2006 的数据对 LPD-1、LID-1、LID-2 以及 LD-1 进行系统的比较，结果发现：(1) LD-1 模型的相对拟合指标最好；(2)在 PCE、TE 和 CCE 的影响中，TE 影响最大，PCE 最小。

综上所述，上述模型都是基于随机效应处理 LD。无论是 LPD-1，还是 LID-1、LID-2，实际上都是通过随机效应处理不同的 LD，这样可以提高 IRT 模型参数估计的准确性(Koziol, 2016)。实际上，也可以通过固定效应处理题组造成的 LID (参见 Hoskens & De Boeck, 1997)。比如，研究者也可以构建类似 3.1 和 3.2 节呈现的三类模型，以系统地讨论 TE 的影响。

首先，这里仅展示基于 Rasch 模型的 EIRTM，实际上 LID 模型可以轻易拓展至两参数逻辑斯蒂克(*two parameter logistic*, 2PL)模型(Fukuhara & Kamata, 2011)，多级记分模型(Jiao & Zhang, 2015)，以及多维模型(Fujimoto, 2018)。其次，不同测量情境可以自由组合，LD-1 是结合 LID 和 LPD 而得到。还可以在 DIF-1 上加入 TE 或 PCE，此类 EIRTM 相比传统 DIF 方法更具有优势(Jin & Kang, 2016; Teker & Dogan, 2015)，甚至可估计题组水平的 DIF (Paek & Fukuhara, 2015; Ravand, 2015)。此外，已有研究基于真实数据进行分析完形填空和阅读理解(Baghaei & Ravand, 2016)。总之，EIRTM 的应用非常灵活，研究者可以基于自身需要与前文提到的 IPE、TME、DIF 模型相结合，构建功能更为强大的模型。

4 实例

此处使用言语攻击数据(Vansteelandt, 2000)对 EIRTM 的使用进行说明。数据包括 316 名学生(73 名男生和 243 名女生)在 24 道题目上的作答。每个题目对应一个情境，由 3 个因素决定：情境类型(本人责任，他人责任)、行为类型(诅咒，责备，怒骂)和行为模式(做，想)。共有 $2\times2\times3 = 12$ 种情境，每种情境有 2 道题。具体如表 1 所示。

将原始的三类作答(“不”、“也许” 以及 “是”)，转换为 0(“不” 与 “也许”)和 1(“是”)评分后，基于 JAGS (Just Another Gibbs Sampler; Plummer, 2017)软件，采用 R 语言“R2jags”包(Su & Yajima, 2015)

调用控制，对此数据进行分析。如需相关代码，可与作者联系。出于解释的方便，所有模型基于 Rasch 模型簇，主要结果如表 2 所示。

模型 1 即为最为基本的 Rasch 模型，对应的 EIRTM 为：

$$\eta_{pi} = \log\left(\frac{P(Y_{pi}=1)}{1-P(Y_{pi}=1)}\right) = \theta_p - \sum_{q=1}^{Q}\beta_q X_{iq} \quad (14)$$

上式中记号的含义与前文一致。以被试 p 在第 1 题上的系统成分为例，$\eta_{p1} = \theta_p - \sum_{q=1}^{Q}\beta_q X_{1q} = \theta_p - \beta_1 = \theta_p - (-1.162)$，易知 β_q 对应各个题目的难度。

模型 2 类似 3.2 中的 TME，这里估计的是行为模式效应。注意模型 2 与 TME 的测验设计有所不同，但是模型是等价的。量表的前 12 道题目是“想”，后 12 题是“做”，这里直接估计出行为模式的效应为−0.465(对应 TME-1 模型)，EIRTM 如下：

$$\eta_{pi} = \log\left(\frac{P(Y_{pi}=1)}{1-P(Y_{pi}=1)}\right) = \theta_p - \sum_{q=1}^{Q}\beta_q X_{iq} + \delta_m M_{\{i>I\}} \quad (15)$$

记号含义与前文一致。被试 p 在第 1 题上的系统成分为：$\eta_{p1} = \theta_p - \beta_1 = \theta_p - (-1.148)$，而被试 p 在第 13 题上的系统成分为：$\eta_{p13} = \theta_p - \beta_{13} + \delta_m = \theta_p - (-1.580) + (-0.465)$。易知 δ_m 对应不同模式造成的效应。

模型 3 对应 3.3 中的 DIF 模型，出于说明的方便，这里没有讨论男女组能力均值不同的情况，对应的 EIRTM 公式如下：

$$\eta_{pi} = \log\left(\frac{P(Y_{pi}=1)}{1-P(Y_{pi}=1)}\right) = \theta_p - \sum_{q=1}^{Q}\beta_q X_{iq} + \sum_{q=1}^{Q}\delta_{ig} X_{iq} Z_g \quad (16)$$

表 1　24 道言语攻击题目

题目	行为模式	情境类型	行为类型
一辆公交车没有进站停靠，我想诅咒。	想	他人责任	诅咒
一辆公交车没有进站停靠，我想责备。	想	他人责任	责备
一辆公交车没有进站停靠，我想怒骂。	想	他人责任	怒骂
因为工作人员给我错误的信息，我错过了火车，我想诅咒。	想	他人责任	诅咒
因为工作人员给我错误的信息，我错过了火车，我想责备。	想	他人责任	责备
因为工作人员给我错误的信息，我错过了火车，我想怒骂。	想	他人责任	怒骂
当我刚进入商店，商店就关门了，我想诅咒。	想	自己责任	诅咒
当我刚进入商店，商店就关门了，我想责备。	想	自己责任	责备
当我刚进入商店，商店就关门了，我想怒骂。	想	自己责任	怒骂
我与对方的通话断了，因为我用完了话费，我想诅咒。	想	自己责任	诅咒
我与对方的通话断了，因为我用完了话费，我想责备。	想	自己责任	责备
我与对方的通话断了，因为我用完了话费，我想怒骂。	想	自己责任	怒骂
一辆公交车没有进站停靠，我会诅咒。	做	他人责任	诅咒
一辆公交车没有进站停靠，我会责备。	做	他人责任	责备
一辆公交车没有进站停靠，我会怒骂。	做	他人责任	怒骂
因为工作人员给我错误的信息，我错过了火车，我会诅咒。	做	他人责任	诅咒
因为工作人员给我错误的信息，我错过了火车，我会责备。	做	他人责任	责备
因为工作人员给我错误的信息，我错过了火车，我会怒骂。	做	他人责任	怒骂
当我刚进入商店，商店就关门了，我会诅咒。	做	自己责任	诅咒
当我刚进入商店，商店就关门了，我会责备。	做	自己责任	责备
当我刚进入商店，商店就关门了，我会怒骂。	做	自己责任	怒骂
我与对方的通话断了，因为我用完了话费，我会诅咒。	做	自己责任	诅咒
我与对方的通话断了，因为我用完了话费，我会责备。	做	自己责任	责备
我与对方的通话断了，因为我用完了话费，我会怒骂。	做	自己责任	怒骂

表 2 24 道言语攻击题目的固定效应

题目	模型 1	模型 2		模型 3			模型 4
	β_q	β_q	行为模式	β_q	DIF	95%置信区间	β_q
1	−1.162	−1.148		−1.196	−0.101	(−0.723, 0.549)	−1.248
2	−0.546	−0.531		−0.574	−0.104	(−0.717, 0.505)	−0.584
3	−0.091	−0.074		−0.134	−0.171	(−0.777, 0.431)	−0.101
4	−1.657	−1.641		−1.727	−0.261	(−0.934, 0.449)	−1.800
5	−0.681	−0.667		−0.729	−0.182	(−0.800, 0.433)	−0.746
6	−0.026	−0.011		−0.184	−0.684	(−1.293, −0.070)	−0.031
7	−0.512	−0.496		−0.495	0.103	(−0.507, 0.721)	−0.617
8	0.630	0.643		0.751	0.535	(−0.067, 1.151)	0.689
9	1.430	1.451		1.338	−0.455	(−1.153, 0.240)	1.610
10	−1.014	−0.998		−1.071	−0.221	(−0.853, 0.415)	−1.221
11	0.312	0.329		0.362	0.231	(−0.376, 0.826)	0.354
12	0.963	0.982		0.866	−0.454	(−1.104, 0.185)	1.132
13	−1.145	−1.580	−0.465	−1.066	0.426	(−0.251, 1.108)	−1.225
14	−0.383	−0.820	−0.465	−0.215	0.792	(0.156, 1.420)	−0.412
15	0.820	0.381	−0.465	0.786	−0.133	(−0.767, 0.487)	0.885
16	−0.822	−1.260	−0.465	−0.618	1.006	(0.352, 1.706)	−0.895
17	0.035	−0.404	−0.465	0.263	1.019	(0.409, 1.648)	0.042
18	1.372	0.933	−0.465	1.422	0.222	(−0.417, 0.879)	1.498
19	0.200	−0.240	−0.465	0.393	0.864	(0.280, 1.481)	0.199
20	1.390	0.956	−0.465	1.579	0.750	(0.093, 1.390)	1.563
21	2.711	2.277	−0.465	2.775	0.244	(−0.615, 1.062)	3.034
22	−0.660	−1.106	−0.465	−0.548	0.568	(−0.068, 1.205)	−0.801
23	0.363	−0.080	−0.465	0.488	0.546	(−0.059, 1.146)	0.416
24	1.867	1.427	−0.465	1.799	−0.359	(−1.138, 0.375)	2.202

这里将女性作为参照组($Z_g=0$)，男性作为目标组($Z_g=1$)。所以女生 p 在题目 1 上的系统成分为：$\eta_{p1}=\theta_p-\beta_1=\theta_p-(-1.196)$，男性 m 在题目 1 上的系统成分为：$\eta_{m1}=\theta_m-\beta_1=\theta_m-(-1.196)+(-0.101)$。$\delta_{ig}$ 对应题目的 DIF 效应量，结合提供的 95%的置信区间，就可以直接判断 δ_{ig} 是否显著。此处，第 6、14、16、17、19、20 题的 DIF 效应显著。

模型 4 考虑的是 3.4 中提到的 CCE，对应的 EIRTM 如下：

$$\eta_{pi}=\log\left(\frac{P(Y_{pi}=1)}{1-P(Y_{pi}=1)}\right)=$$
$$\theta_p-\sum_{q=1}^{Q}\beta_q X_{iq}+\sum_{c=1}^{C}\gamma'_{pc}T'_{ic} \quad (17)$$

由表 4 易知量表的内容(题干)能够归为 4 类，对应 4 个随机效应 γ'_{pc}。不同被试在不同内容上的 γ'_{pc} 都不同，以第 1 个内容为例，$\gamma'_{p1}\sim N(0.004,0.442)$。当具体到被试 1 在题目 1 上的作答时，JAGS 可以估计出 γ'_{11} 的值为−0.398，系统成分为：$\eta_{11}=\theta_1-\beta_1+\gamma'_{11}=\theta_1-(-1.248)+(-0.398)$；被试 1 在题目 2 上作答时，由于属于同一个内容，系统成分为：$\eta_{12}=\theta_1-\beta_2+\gamma'_{11}=\theta_1-(-0.584)+(-0.398)$。

最后，值得一提的是 JAGS 采用的是贝叶斯方法，可以通过离差信息指数(*Deviance Information Criterion*，DIC)来评估模型的整体拟合情况，DIC 越小说明模型的预测能力越好。这 4 个模型中，模型 3 的 DIC 最小(DIC = 7855.3)，即拟合最好。

5 讨论与展望

5.1 将 EIRTM 用于测量不变性研究

本文的第 3 部分详细介绍了如何使用 EIRTM 检测 IPE、TME 以及 DIF，这些都反映 EIRTM 能够方便地处理测量不变性问题：IPE 是题目位置

对测量不变性的影响, TME 是测验形式对测量不变性的影响, DIF 是受测群体对测量不变性的影响。通过 EIRTM 处理测量不变性问题可以解决传统 IRT 方法(即"两步法")的困境：如果测量不变性不满足, 那么 IRT 得到的参数估计本身就是有偏的; 基于有偏的参数估计, 并不能得到可信的结果。因此即使基于"两步法"证明数据满足测量不变性, 也有可能是不准确的参数估计造成的。

此外, EIRTM 可以构建全面的测量不变性模型, 得到尽可能准确的参数估计结果。读者可能已经意识到, 鉴于 EIRTM 的灵活性, 可以将第 3 部分中提到的模型进行整合, 得到一个既能估计 IPE、TME 和 DIF, 又考虑 LD 的模型。换言之, 只要符合研究实际, 研究者可以一步到位, 同时处理多个测量问题。

最后, EIRTM 可以将测量不变性问题与解释性分析相结合, 也即在估计 IPE、TME 或 DIF 的同时, 也考虑被试和题目特征的影响。此类模型能够通过控制测量不变性的相关效应, 得到更为准确的被试和题目效应; 反之亦然。实际上, DIF-1 就是在控制组别的固定效应后, 再估计 DIF 效应。

5.2　通过 EIRTM 构建综合性的分析框架

EIRTM 提供一个统一而灵活的 IRT 模型框架, 并且越来越受到研究者重视。受限于篇幅和主旨, 本文没法更全面地展示 EIRTM 与现有 IRT 模型的转换关系, 除本文涉及的模型外, 使用 EIRTM 还可以建构多级记分的 IRT 模型和多维 IRT 模型、动态 Rasch 模型(*Dynamic Rasch Models*)、纵向 IRT 模型以及含反应时的 IRT 模型等等(参见 De Boeck & Wilson, 2004; Klein Entink, Kuhn, Hornke, & Fox, 2009; Rijmen et al., 2003; Wilson, Zheng, & McGuire, 2012)。以 EIRTM 为代表的广义建模方法(*Generalized Modeling Approaches*)具有诸多优越性, 目前已经得到业内研究者的重视。在新编著的《项目反应理论手册(第一卷)：模型》(*Handbook of Item Response Theory, Volume One: Models*; van der Linden, 2016)的最后一部分, 专门介绍了 4 种广义建模方法, 这值得国内研究者重视。

此外, EIRTM 还体现了 IRT 模型和回归模型的共性。传统的心理和教育测量领域中, 很少有研究者注意到回归模型、GLM、HLM 和 IRT 模型之间的联系：在回归模型的基础上, 加入随机效应, 可以推广至 HLM; 引入连接函数, 可以得到 GLM; 同时加入随机效应和连接函数, 可以得到 EIRTM。这一综合的分析框架, 不仅有助于研究者深入认识以 IRT 为代表的现代测量理论与经典回归分析的联系, 也有利于相应的教学和实践活动。

5.3　EIRTM 的应用前景与不足

EIRTM 具有广阔的应用前景, 可以广泛应用于心理和教育测量领域中。除了上文所述的通过 EIRTM 建构合理的测量模型以外, EIRTM 还可用于分析复杂表现任务(*complex performance task*)。对于复杂表现任务进行评价, 是教育与心理测量领域面临的新挑战(Mislevy, 2016)。比如, PISA 2015 就使用合作问题解决任务, 以展示学生在动态、交互情景中的表现(OECD, 2017b)。EIRTM 以其灵活的框架为评价复杂表现任务提供了一种解决思路, 通过 EIRTM 可以将涉及的任务属性的特征纳入模型, 从而得到被试能力的准确估计。

当然 EIRTM 也存在一些问题：(1) 算法比较复杂, 运算时间相对较长。对于蒙特卡洛(*Monte Carlo*)模拟研究以及自适应测验而言, 只能尝试通过提高计算机的计算性能来改进效率。但是对于不需要重复的应用研究来说, 现有软件的运行速度基本可以接受; (2) EIRTM 的使用对数学能力和编程能力要求较高, 这不太利于一般研究者的使用。EIRTM 涉及的算法比较复杂, 非统计学/数学专业的研究者不容易理解; 而且目前没有简单易用的专用软件可供使用, 必须由研究者自己编写程序, 并设定模型参数。总之, 尽管 EIRTM 也存在一些不足, 但是考虑到 EIRTM 的重要理论意义与应用价值, 未来必定能在测量领域大有作为。

致谢：感谢美国罗格斯大学心理测量专业在读博士孙研对本文的英文摘要进行修改和润色, 感谢北京师范大学中国基础教育质量监测协同创新中心的薛明峰同学和统计学院的任赫同学对文章内容的修正。

参考文献

刘红云, 骆方. (2008). 多水平项目反应理论模型在测验发展中的应用. *心理学报*, *40*(1), 92–100.

聂旭刚, 陈平, 张缨斌, 何引红. (2018). 题目位置效应的概念及检测. *心理科学进展*, *26*(2), 368–380.

詹沛达, 王文中, 王立君. (2013). 项目反应理论新进展之题组反应理论. *心理科学进展*, *21*(12), 2265–2280.

Adams, R. J., Wu, M. L., & Wilson, M. R. (1988). *ACER ConQuest: Generalised item response modelling software* [Computer software]. Melbourne, Victoria, Australia: Australian Council for Educational Research.

Baghaei, P., Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicologica: International Journal of Methodology and Experimental Psychology, 37*(1), 85–104.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C (2015). Fitting linear mixed-effects models using LME4. *Journal of Statistical Software, 67*(1), 1–48.

Bechger, T. M., Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika, 80*(2), 317–340.

Binet, A., & Simon, T. (1904). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'année Psychologique*, *11*(1), 191–244.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, 35(2), 179–197.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*(2), 113–141.

Cosgrove, J., & Cartwright, F. (2014). Changes in achievement on PISA: The case of Ireland and implications for international assessment practice. *Large Scale Assessments in Education, 2*(2), 1–17.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164–185.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39*(6), 502–523.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1–28.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

De Boeck, P., Wilson, M. R. (2016). Explanatory response models. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory, Volume One: Models* (pp. 565–580). New York, NY: Chapman and Hall/CRC.

Eyre, J., Berg, M., Mazengarb, J., & Lawes, E. (2017). Mode equivalency in PAT: Reading comprehension. Wellington: NZCER.

Fujimoto, K. A. (2018). A general Bayesian multilevel multidimensional IRT model for locally dependent data. *British Journal of Mathematical and Statistical Psychology, 71*(3), 536–560.

Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement, 35*(8), 604–622.

Gamerman, D., Gonçalves, F. B., Soares, T. M. (2018). Differential item functioning. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory, Volume Three: Applications* (pp. 67–86). New York, NY: Chapman and Hall/CRC.

Gill, J. (2000). *Generalized linear models: A unified approach* (Vol. 134). Thousand Oaks, CA: Sage Publications.

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling, 54*(4), 418–431.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analyzing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation, 17*(6), 497–509.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, *2*(3), 261–277.

Ip, E. H. (2000). Adjusting for information inflation due to local dependency in moderately large item clusters. *Psychometrika, 65*(1), 73–91.

Janssen, R. (2016). Linear Logistic Models. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory, Volume One: Models* (pp. 211–224). New York, NY: Chapman and Hall/CRC.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics, 38*(1), 32–60.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2014). Flexible item response theory modeling with FLIRT. *Applied Psychological Measurement*, *38*(5), 404–405.

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education:*

Principles, Policy & Practice, 23(4), 495–518.

Jerrim, J., Micklewright, J., Heine, J. H., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education*, *44*(4), 476–493.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*(1), 82–100.

Jiao, H., Kamata, A., & Xie, C. (2015). Multilevel cross-classified testlet model for complex item and person clustering in item response data analysis. In J. R. Harring, L. M. Stapleton & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 139–161). Charlotte, NC: Information Age Publishing Inc.

Jiao, H., Wang, S. D., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement, 6*(3), 311–321.

Jiao, H., Zhang, Y. (2015). Polytomous multilevel testlet models for testlet-based assessments with complex sampling designs. *British Journal of Mathematical and Statistical Psychology, 68*(1), 65–83.

Jin, Y., Kang, M. (2016). Comparing DIF methods for data with dual dependency. *Large-scale Assessments in Education, 4*(1), 18.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79–93.

Kang, C. (2014). *Linear and nonlinear modeling of item position effects* (Unpublished master's thesis). University of Nebraska-Lincoln.

Klein Entink, R. H., Kuhn, J. T., Hornke, L. F., & Fox, J. P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological methods*, *14*(1), 54–75.

Koziol, N. A. (2016). Parameter recovery and classification accuracy under conditions of testlet dependency: A comparison of the traditional 2PL, testlet, and bi-factor models. *Applied Measurement in Education, 29*(3), 184–195.

Lee, Y. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing, 21*(1), 74–100.

Logan, T. (2015). The influence of test mode and visuospatial ability on mathematics assessment performance. *Mathematics Education Research Journal*, 27(4), 423–441.

Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement*, *53*(3), 265–292.

OECD. (2017a). *PISA 2015 technical report*. Pairs: OECD Publishing.

OECD. (2017b). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving*, Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264281820-en.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Vol. 161). Thousand Oaks, CA: Sage Publications.

Paek, I., Fukuhara, H. (2015). Estimating a DIF decomposition model using a random-weights linear logistic test model approach. *Behavior Research Methods, 47*(3), 890–901.

Plummer, M. (2017). *JAGS version 4.3.0 user manual* [Software manual]. Retrieved from https://martynplummer.wordpress.com/2017/07/18/jags-4-3-0-is-released/

Rabe-Hesketh, S., Skrondal, A. (2016). Generalized linear latent and mixed modeling. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory, Volume One: Models* (pp. 503–526). New York, NY: Chapman and Hall/CRC.

Rabe-Hesketh, S., Skrondal, A & Pickles, A. (2004). *GLLAMM manual* [Software manual]. (U. C. Berkeley Division of Biostatistics Working Paper Series, 160)

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon Jr, R. T., & Toit, M. D. (2011). *HLM7 hierarchical linear and nonlinear modeling manual* [Software manual]. Lincolnwood, IL: SSI Scientific Software International Inc.

Ravand, H. (2015). Assessing testlet effect, impact, differential testlet, and item functioning using cross-classified multilevel measurement modeling. *SAGE Open, 5*(2).

Rijmen, F. (2006). *BNL: A Matlab toolbox for Bayesian networks with logistic regression* (Tech. Rep.). Amsterdam, the Netherlands: VU University Medical Center.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*(2), 185–205.

SAS Institute. (2015). *SAS/STAT 14.1: user's guide* [Software manual]. Cary, NC: SAS Institute Inc.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2014). *OpenBUGS (Version 3.2.3)* [Software manual]. Retrieved from, http://www.openbugs.net/Manuals/Manual.html.

Stroup, W. W. (2012). *Generalized linear mixed models: Modern concepts, methods and applications*. Boca Raton, FL: CRC press.

Su Y, Yajima M (2015). *R2jags: A Package for Running JAGS from R* [Computer software]. Retrieved from http://CRAN.R-project.org/package=R2jags.

Teker, G. T., Dogan, N. (2015). The Effects of testlets on reliability and differential item functioning. *Educational Sciences: Theory and Practice, 15*(4), 969–980.

Thissen, D. (1991). *MULTILOG* [Software manual]. Lincolnwood, IL: Scientific Software.

Trendtel, M., Robitzsch, A. (2018). Modeling item position

effects with a Bayesian item response model applied to PISA 2009–2015 data. *Psychological Test and Assessment Modeling, 60*(2), 241–263.

Tutz, G., Berger, M. (2016). Item-focussed trees for the identification of items in differential item functioning. *Psychometrika, 81*(3), 727–750.

Tutz, G., Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika, 80*(1), 21–43.

van der Linden, W. J. (2016). *Handbook of Item Response Theory, Volume One*. New York, NY: Chapman and Hall/CRC.

van der Linden, W. J. (2018). *Handbook of Item Response Theory, Volume Three: Applications*. New York, NY: Chapman and Hall/CRC.

Vansteelandt, K. (2000). *Formal models for contextualized personality psychology* (Unpublished doctoral dissertation). K.U. Leuven, Belgium.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57*(5), 741–758.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). *Differential testlet functioning definitions and detection* (Research Rep. 91-21). Princeton NJ: ETS.

Wang, W. C., & Wilson, M. (2005). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement*, *65*(4), 549–576.

Weirich, S., Hecht, M., Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38*(7), 535–548.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied psychological measurement, 41*(2), 115–129.

Wilson, M., Zheng, X. H., & McGuire, L. (2012). Formulating latent growth using an explanatory item response model approach. *Journal of Applied Measurement*, *13*(1), 1–22.

Xie, C. (2014). *Cross-classified modeling of dual local item dependence* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Xie, C., & Jiao, H. (2014, April). *Cross-classified modeling of dual local item dependence*. Paper presented at the Annual Meeting of the American Educational Research Association, Phliadelphia, PA.

Explanatory item response theory models: Theory and application

CHEN Guanyu; CHEN Ping

(*Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing* 100875, *China*)

Abstract: Explanatory item response theory models (EIRTM) refer to a family of item response theory (IRT) models that are constructed based on the generalized linear mixed models and nonlinear mixed models. EIRTM can be utilized to address various measurement problems by incorporating predictors into IRT models. First, the relevant concepts and parameter estimation methods of EIRTM are introduced in this paper, followed by the procedures regarding how to use EIRTM to account for the item position effect, test mode effect, differential item functioning, local person dependence, and local item dependence. Next, an example is provided to illustrate the use of EIRTM. Finally, the shortcomings and potential applications of EIRTM are discussed.

Key words: explanatory item response theory; generalized linear mixed models; nonlinear mixed models; measurement invariance; explanatory measurement